A Proofs

A.1 Proof of Proposition 1

For simplicity of notation, we will drop the hats from all of the variables in this proof, since they are irrelevant to the result. Without loss of generality, we can also assume $Y := \{a_1, \dots, a_{|Y|}\} = \{1, \dots, K\}$. To show that $B_{:Y}(L_Y)^{-1}(B_{:Y})^\top = \sum_{k=1}^{|Y|} \boldsymbol{c}_k \boldsymbol{c}_k^\top$, we proceed by induction on K. When $K = 1$, the result is trivial since:

$$B_{:Y}(L_Y)^{-1}(B_{:Y})^\top = \boldsymbol{b}_1(\boldsymbol{b}_1^\top \boldsymbol{b}_1)^{-1}\boldsymbol{b}_1^\top = \frac{\boldsymbol{b}_1 \boldsymbol{b}_1^\top}{\|b_1\|_2^2} = \boldsymbol{c}_1 \boldsymbol{c}_1^\top. \tag{21}$$

For the inductive case, we assume that the statement holds for $K-1$. For simplicity, we denote $X := [\boldsymbol{b}_1, \dots, \boldsymbol{b}_{K-1}]$ and $C_k := \sum_{j=1}^{k} \boldsymbol{c}_j \boldsymbol{c}_j^\top$ for $k = 1, \dots, K$. From the inductive hypothesis, it holds that:

$$X(X^\top X)^{-1}X^\top = C_{K-1} = \sum_{j=1}^{K-1} \boldsymbol{c}_j \boldsymbol{c}_j^\top. \tag{22}$$

Writing the $D \times K$ matrix $B_{:Y}$ in terms of its submatrix X, we have:

$$\begin{aligned} B_{:Y}(L_Y)^{-1}B_{:Y}^\top = B_{:Y}(B_{:Y}^\top B_{:Y})^{-1}B_{:Y}^\top &= \begin{bmatrix} X & \boldsymbol{b}_K \end{bmatrix} \left(\begin{bmatrix} X^\top \\ \boldsymbol{b}_K^\top \end{bmatrix} \begin{bmatrix} X & \boldsymbol{b}_K \end{bmatrix} \right)^{-1} \begin{bmatrix} X^\top \\ \boldsymbol{b}_K^\top \end{bmatrix} \\ &= \begin{bmatrix} X & \boldsymbol{b}_K \end{bmatrix} \begin{bmatrix} X^\top X & X^\top \boldsymbol{b}_K \\ \boldsymbol{b}_K^\top X & \boldsymbol{b}_K^\top \boldsymbol{b}_K \end{bmatrix}^{-1} \begin{bmatrix} X^\top \\ \boldsymbol{b}_K^\top \end{bmatrix}. \end{aligned} \tag{23}$$

From the Schur complement, we get that:

$$\begin{aligned} \begin{bmatrix} X^\top X & X^\top \boldsymbol{b}_K \\ \boldsymbol{b}_K^\top X & \boldsymbol{b}_K^\top \boldsymbol{b}_K \end{bmatrix}^{-1} = \begin{bmatrix} (X^\top X)^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} + & \\ \frac{1}{\boldsymbol{b}_K^\top(\boldsymbol{b}_K - X(X^\top X)^{-1}X^\top \boldsymbol{b}_K)} & \begin{bmatrix} (X^\top X)^{-1}X^\top \boldsymbol{b}_K \boldsymbol{b}_K^\top X (X^\top X)^{-1} & -(X^\top X)^{-1}X^\top \boldsymbol{b}_K \\ -\boldsymbol{b}_K^\top X (X^\top X)^{-1} & 1 \end{bmatrix} \end{aligned} \tag{24}$$

where $\boldsymbol{0}$ is k-dimensional zeros vector. Substituting Equation 24 into Equation 23, we obtain:

$$\begin{aligned} B_{:Y}(L_Y)^{-1}B_{:Y}^\top &= X(X^\top X)^{-1}X^\top + \frac{1}{\boldsymbol{b}_K^\top(\boldsymbol{b}_K - C_{K-1}\boldsymbol{b}_K)} \left(C_{K-1}\boldsymbol{b}_K \boldsymbol{b}_K^\top C_{K-1} - C_{K-1}\boldsymbol{b}_K \boldsymbol{b}_K^\top - \boldsymbol{b}_K \boldsymbol{b}_K^\top C_{K-1} + \boldsymbol{b}_K \boldsymbol{b}_K^\top \right) \\ &= C_{K-1} + \frac{(\boldsymbol{b}_K - C_{K-1}\boldsymbol{b}_K)(\boldsymbol{b}_K - C_{K-1}\boldsymbol{b}_K)^\top}{\boldsymbol{b}_K^\top(\boldsymbol{b}_K - C_{K-1}\boldsymbol{b}_K)} \\ &= C_{K-1} + \frac{\boldsymbol{d}_K \boldsymbol{d}_K^\top}{\boldsymbol{b}_K \boldsymbol{d}_K} \\ &= C_{K-1} + \boldsymbol{c}_K \boldsymbol{c}_K^\top = \sum_{j=1}^{K} \boldsymbol{c}_j \boldsymbol{c}_j^\top. \end{aligned} \tag{25}$$

This completes the proof of Proposition 1.

A.2 Proof of Theorem 1

First, recall our assumption that the smallest singular value of $\widehat{B}_{:S}$ is greater than 1 for all $S \subseteq [N]$ where $|S| = K$. This implies that the eigenvalues of $\widehat{L}_S = \widehat{B}_{:S}^\top \widehat{B}_{:S}$ are all greater than 1. It also implies, by the eigenvalue interlacing theorem, that the eigenvalues of all submatrices of this matrix are greater than 1. This means that, for any $S \subseteq [N]$ where $|S| \leq K$, the determinant of the matrix $\widehat{L}_S$ is always > 1. This is sufficient to show that the set function $f(S) = \log\det(\widehat{L}_S)$ defined on $S \subseteq [N], |S| \leq K$ is non-negative, monotone, and submodular (see Sharma et al., 2015, Propositions 1 and 2). These three properties, combined with the fact that

$\log(\det(\widehat{L}_\emptyset)) = \log(1) = 0$, imply that the greedy algorithm of Nemhauser et al. (1978) applied to the log det function gives a $(1 - 1/e)$-approximation ratio:

$$\log\det(\widehat{L}_{\bar{Y}}) \geq (1 - 1/e)\log\det(\widehat{L}_{Y^*}). \tag{26}$$

By monotonicity of log, we get the statement from the theorem. This completes the proof of Theorem 1.

A.3 Proof of Theorem 2

From our stated minumum eigenvalue assumption on $\widehat{B}_{:S}$, we have that $f(S) = \log\det(\widehat{L}_S)$ defined on $S \subseteq [N], |S| \leq K$ is non-negative, monotone, and submodular. (See the proof of Theorem 1 for a more in-depth argument of why this is the case.) The properties guarantee that the greedy algorithm of Nemhauser et al. (1978) applied to the log det function gives a $(1 - 1/e)$-approximation ratio. The greedy algorithm starts from the empty set and builds a set S one item at a time by adding the item that maximizes the following marginal gain:

$$\operatorname*{argmax}_{i\in[N]} \log\left(\frac{\det(\widehat{L}_{S\cup\{i\}})}{\det(\widehat{L}_S)}\right). \tag{27}$$

Recall from the developments in Equation 4 through Equation 11 that we can write this marginal gain as the following inner product:

$$\log\left(\frac{\det(\widehat{L}_{S\cup\{i\}})}{\det(\widehat{L}_S)}\right) = \log\left\langle W^\top W - W^\top \widehat{C}^{(S)} W, \boldsymbol{b}_i \boldsymbol{b}_i^\top \right\rangle. \tag{28}$$

Further, note that, since we already established this log det is non-negative, the inner product here is a positive quantity (in fact, it must be ≥ 1). This means that, if $i \in [N]$ is the index from a $(1-\varepsilon)$-approximate MIPS structure $\mathcal{M}$ with the query matrix $W^\top W - W^\top \widehat{C}^{(S)} W$, we have the following lower bound:

$$\begin{aligned}\log\frac{\det(\widehat{L}_{S\cup\{i\}})}{\det(\widehat{L}_S)} &\geq \log\left((1-\varepsilon)\max_{i\in[N]}\left\langle W^\top W - W^\top \widehat{C}^{(S)} W, \boldsymbol{b}_i \boldsymbol{b}_i^\top \right\rangle\right) \\ &= -\log\left(\frac{1}{1-\varepsilon}\right) + \max_{i\in[N]}\log\frac{\det(\widehat{L}_{S\cup\{i\}})}{\det(\widehat{L}_S)}. \end{aligned} \tag{29}$$

To conclude the proof, we introduce the following lemma that guarantees the optimality of the greedy algorithm when the marginal gain is maximized with an additive error.

Lemma 1. *Suppose that $f : [N] \to \mathbb{R}_+$ is monotone submodular function that satisfies $f(\emptyset) = 0$. Let $Y_0 = \emptyset$ and $Y_i = Y_{i-1} \cup \{a_i\}$ such that:*

$$f(Y_{i-1} \cup \{a_i\}) - f(Y_{i-1}) \geq \max_{a\in[N]} f(Y_{i-1} \cup \{a\}) - f(Y_{i-1}) - \alpha \tag{30}$$

for $i = 1, \ldots, K$ and some $\alpha \geq 0$. Then, it holds that:

$$f(Y_K) \geq \left(1 - \frac{1}{e}\right)\max_{|S|=K} f(S) - K\alpha. \tag{31}$$

From Lemma 1 and Equation 29, we have that:

$$\log\det(\widehat{L}_Y) \geq \left(1 - \frac{1}{e}\right)\max_{|S|=K}\log\det(\widehat{L}_S) - K\log\left(\frac{1}{1-\varepsilon}\right), \tag{32}$$

where Y is the output of Algorithm 1. Exponentiating, we obtain the result:

$$\det(\widehat{L}_Y) \geq (1-\varepsilon)^K\left(\max_{|S|=K}\det(\widehat{L}_S)\right)^{(1-1/e)}. \tag{33}$$

This concludes the proof of Theorem 2.

A.4 Proof of Lemma 1

This proof is similar to that of Nemhauser et al. (1978)'s proof of the $(1-1/e)$-approximation guarantee for the greedy algorithm, with a few modifications. Let $Y_i = \{a_1, a_2, \dots, a_i\}$ for $i = 1, \dots, K$ and let $Y^* := \operatorname{argmax}_{|S|=K} \det(\widehat{L}_S)$ represent the optimal size-K set. For a fixed i, let $Y^* \setminus Y_i := \{d_1, \dots, d_r\}$. Then:

$$\begin{aligned}
f(Y^*) &\le f(Y^* \cup Y_i) && (34)\\
&= f(Y_i) + [f(Y_i \cup \{d_1\}) - f(Y_i)] + [f(Y_i \cup \{d_1, d_2\}) - f(Y_i \cup \{d_1\})] + && (35)\\
&\quad \cdots + [f(Y_i \cup Y^*) - f(Y_i \cup (Y^* \setminus d_r))] && (36)\\
&\le f(Y_i) + |Y^* \setminus Y_i| \left(\max_{\ell \in [r]} f(Y_i \cup \{d_\ell\}) - f(Y_i) \right) && (37)\\
&\le f(Y_i) + K \left(\max_{\ell \in [r]} f(Y_i \cup \{d_\ell\}) - f(Y_i) \right) && (38)\\
&\le f(Y_i) + K \left(\max_{j \in [N]} f(Y_i \cup \{j\}) - f(Y_i) \right) && (39)\\
&\le f(Y_i) + K \left(f(Y_i \cup \{a_{i+1}\}) - f(Y_i) + \alpha \right), && (40)
\end{aligned}$$

where the first inequality follows from monotonicity of f, the second from the submodularity of f, the third from the fact that $|Y^*| \le K$, and the last from the greedy selection property. Rearranging the above equation, we have that:

$$f(Y^*) - f(Y_{i+1}) \le \left(1 - \frac{1}{K}\right) (f(Y^*) - f(Y_i)) + \alpha. \tag{41}$$

Using induction on i, it holds that:

$$\begin{aligned}
f(Y^*) - f(Y_K) &\le \left(1 - \frac{1}{K}\right)^K (f(Y^*) - f(Y_0)) + \alpha \sum_{i=0}^{K} \left(1 - \frac{1}{K}\right)^i && (42)\\
&\le \left(1 - \frac{1}{K}\right)^K f(Y^*) + \alpha K && (43)\\
&\le \frac{1}{e} f(Y^*) + \alpha K. && (44)
\end{aligned}$$

Rearranging this, we conclude the proof of Lemma 1.

A.5 Proof of Theorem 3

From the definition of $\overline{\boldsymbol{b}}_{\mathcal{C}} := \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \boldsymbol{b}_i$ and $\overline{B}_{\mathcal{C}} := \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \boldsymbol{b}_i \boldsymbol{b}_i^\top$, we have that:

$$\begin{aligned}
\overline{\boldsymbol{b}}_{\mathcal{C}} \overline{\boldsymbol{b}}_{\mathcal{C}}^\top - \overline{B}_{\mathcal{C}} &= \frac{\left(\sum_{i \in \mathcal{C}} \boldsymbol{b}_i\right) \left(\sum_{i \in \mathcal{C}} \boldsymbol{b}_i\right)^\top - |\mathcal{C}| \left(\sum_{i \in \mathcal{C}} \boldsymbol{b}_i \boldsymbol{b}_i^\top\right)}{|\mathcal{C}|^2} && (45)\\
&= \frac{\left(\sum_{i \neq j} \boldsymbol{b}_i \boldsymbol{b}_j^\top\right) - (|\mathcal{C}| - 1) \left(\sum_{i \in \mathcal{C}} \boldsymbol{b}_i \boldsymbol{b}_i^\top\right)}{|\mathcal{C}|^2} && (46)\\
&= -\frac{1}{|\mathcal{C}|^2} \sum_{i<j} (\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^\top && (47)
\end{aligned}$$

Putting together this, we have that, for a symmetric query matrix Q:

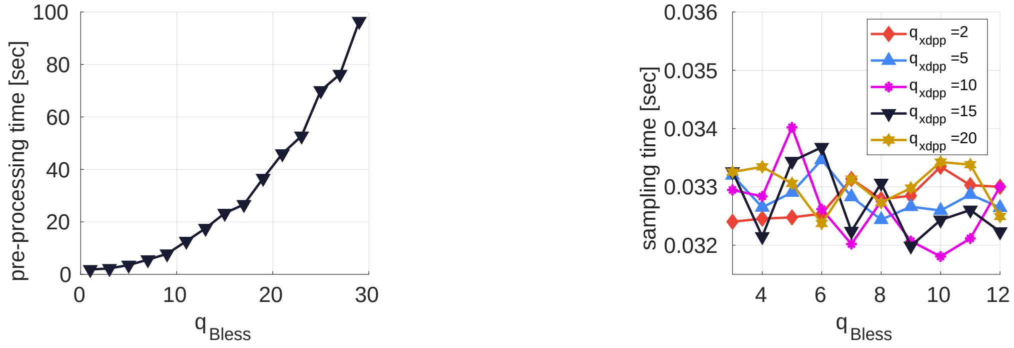
$$\begin{aligned}
\left|\left\langle \bar{\boldsymbol{b}}_{\mathcal{C}}\bar{\boldsymbol{b}}_{\mathcal{C}}^\top, Q\right\rangle - \left\langle \overline{B}_{\mathcal{C}}, Q\right\rangle\right| &= \left|\left\langle \bar{\boldsymbol{b}}_{\mathcal{C}}\bar{\boldsymbol{b}}_{\mathcal{C}}^\top - \overline{B}_{\mathcal{C}}, Q\right\rangle\right| && (48)\\
&= \frac{1}{|\mathcal{C}|^2}\left|\sum_{i<j}\left\langle \left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)\left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)^\top, Q\right\rangle\right| && (49)\\
&\le \frac{1}{|\mathcal{C}|^2}\sum_{i<j}\left|\left\langle \left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)\left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)^\top, Q\right\rangle\right| && (50)\\
&= \frac{1}{|\mathcal{C}|^2}\sum_{i<j}\left|\left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)^\top Q\left(\boldsymbol{b}_i-\boldsymbol{b}_j\right)\right| && (51)\\
&\le \frac{1}{|\mathcal{C}|^2}\left(\sum_{i<j}\left\|\boldsymbol{b}_i-\boldsymbol{b}_j\right\|_2^2\right)\|Q\|_2 && (52)\\
&\le \frac{1}{2}\left(\max_{i,j\in\mathcal{C}}\left\|\boldsymbol{b}_i-\boldsymbol{b}_j\right\|_2^2\right)\|Q\|_2\,. && (53)
\end{aligned}$$

The inequality on the third line follows from the triangle inequality. The equality on the fourth line is from the fact that $\left\langle \boldsymbol{v}\boldsymbol{v}^\top, A\right\rangle = \boldsymbol{v}^\top A\boldsymbol{v}$ for a vector $\boldsymbol{v}$ and a symmetric matrix A. The next-to-last inequality follows from the definition of the 2-norm of a matrix: $\|Q\|_2 = \max_{\|\boldsymbol{x}\|_2=1}\boldsymbol{x}^\top Q\boldsymbol{x}$. This completes the proof of Theorem 3.

B Parameter Search for FastSamp

Even though it produces exact samples from the DPP, FastSamp (Derezinski et al., 2019) has two tunable parameters that have some influence on the runtime of the method: q_{Bless} and q_{xdpp}. We benchmarked the performance of the method with respect to these parameters on synthetic data with $N = 10{,}000$.

The pre-processing time only depends on q_{Bless} (i.e., it is $\mathcal{O}(q_{\text{Bless}}^2)$), and the dependency is shown in the leftmost plot of Figure 3. Item selection (sampling) time depends on both parameters in theory, but in practice we found that the time was in fact very insensitive to how these parameters were set; see the rightmost plot in Figure 3. We choose $q_{\text{Bless}} = 3$ and $q_{\text{xdpp}} = 2$ for our experiments in Section 4. (Setting $\boldsymbol{q}_{\text{Bless}}$ any smaller makes the algorithm unstable.)



(a)

(b)

Figure 3: Results for (a) pre-processing time varying q_{Bless} and (b) item selection (sampling) time varying both q_{Bless} and q_{xdpp}.